Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Agent Deep Reinforcement Learning-Based Fairness-Aware Power Allocation in NOMA Systems under Imperfect SIC for 5G Communication Networks

Mohammed Dahiru Buhari

buharim@kiu.ac.ug

Kampala International University

Bashir Olaniyi Sadiq

Kampala International University

Ogenyi Fabian Chukwudi

Kampala International University

Abdullahi Bala Kunya

Kampala International University

Enerst Edozie

Kampala International University

Research Article

Keywords: Deep Reinforcement Learning (DRL), Deep Q-Network (RL-DQN), Multi-Agent Deep Reinforcement Learning (MARL), NOMA, Quality of service, PSO, Wireless communication

Posted Date: October 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10108243/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multi-Agent Deep Reinforcement Learning-Based Fairness-Aware Power Allocation in NOMA Systems under Imperfect SIC for 5G Communication Networks

Mohammed Dahiru Buhari[1,2], Bashir Olaniyi Sadiq[1,3], Ogenyi Fabian Chukwudi[1], Abdullahi Bala Kunya[1,4], and Enerst Edozie[1]

[1] Department of Electrical and Computer Engineering, Kampala International University, Kampala 20000, Uganda
[2] Department of Electrical and Electronics Engineering, Abubakar Tafawa Balewa University, Bauchi, 740004, Nigeria
[3] Department of Computer Engineering, Ahmadu Bello University, Zaria 610001, Nigeria
[4] Department of Electrical Engineering, Ahmadu Bello University, Zaria 610001, Nigeria
Correspondence: buharim@kiu.ac.ug; dbmohammed@atbu.edu.ng

Abstract

In recent years, Non-Orthogonal Multiple Access (NOMA) has been presented as a key enabler to achieve very high spectral efficiency in conventional 5G networks since it allows multiple users to simultaneously share the same frequency resources by way of power-domain multiplexing. However, efficient power allocation in NOMA remains a challenging task due to the coupled effects of user interference, heterogeneous quality-of-service (QoS) requirements, and the practical limitations of imperfect successive interference cancellation (SIC). Conventional power allocation methodologies usually follow fixed, or equal allocation and therefore cannot dynamically adapt, tending to reduce the fairness and reliability experienced by the users under bad channel conditions. To address these challenges, this paper proposes a Multi-Agent Deep Reinforcement Learning (MARL)-based power allocation framework for downlink NOMA systems operating under imperfect SIC conditions. The proposed MARL framework, unlike traditional centralized methods, takes a user-centered approach of modeling each user as an intelligent agent that will learn the optimal methods of allocating power based on local observations independently. This decentralized formulation is scalable and provides adaptive decision-making in dynamic wireless conditions. In a bid to balance the system performance goals, a multi-objective rewarding functionality is created to be cooperative such that the system sum rate, user fairness, and outage probability are jointly optimized. The effect of residual interference brought about by imperfect SIC is explicitly factored in the learning environment and the agents are then in a position to come up with strong and effective allocation policies. Simulation findings indicate that the proposed MARL framework can attain competitive spectral efficiency, and ensure better scalability, fairness, and performance in different channel conditions than the fixed, equal, Genetic Algorithm (GA), and Reinforcement Learning-Deep Q-Network (RL-DQN) algorithms. At the SNR of 30 dB, the MARL achieved a sum rate of about 18 bps/Hz and a fairness index of 0.96. In addition, it reduces the outage probability (below 10% at 10 dB) and minimizes BER to less than $\mathbf{10^{-3}}$ at high SNR. These findings highlight the potential of decentralized learning-based resource allocation as a viable solution for intelligent and scalable radio resource management in next-generation wireless communication systems.

Keywords: Deep Reinforcement Learning (DRL), Deep Q-Network (RL-DQN), Multi-Agent Deep Reinforcement Learning (MARL), NOMA, Quality of service, PSO, Wireless communication.

1 Introduction

With the explosive growth in wireless data traffic and the increasing demand for high-speed, low-latency, and energy-efficient communications, the fifth-generation (5G) and upcoming sixth-generation (6G) wireless systems have necessitated a fundamental shift from traditional multiple access schemes to more advanced technologies [1]. One very promising innovation is Non-Orthogonal Multiple Access (NOMA). In contrast to traditional Orthogonal Multiple Access (OMA) approaches such as Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), and Orthogonal Frequency Division Multiple Access (OFDMA) in which users are served in orthogonal time, frequency, or code domains, NOMA exploits the power domain to simultaneously serve multiple users over the same resource block [2-3]. This, thus, leads to improvements of spectral efficiency, user fairness, and overall system capacity. These improvements are required for massive connectivity in 5G and beyond. As seen in Fig. 1, in a NOMA system, there is power-domain superposition of users at the transmitter, while SIC is used to separate them at the receiver [4]. As seen here, it is possible for better-channel users to first decode and subtract the signals of the weaker-channel users before decoding their own data. But a NOMA system's performance depends seriously

on the power allocation strategy adopted. The optimal power allocation strategy ensures that users who are in poor channel conditions are given enough power to satisfy their quality of service (QoS) requirements while maximizing the overall throughput of the system [5].

Power allocation in NOMA is an intricate optimization problem, as it is contingent on several parameters such as user channel conditions, target data rates, total available power, interference faced, and the fairness considerations among users [6]. Several power allocations schemes have been proposed in the literature that range from simple heuristic-based approaches such as fixed and equal power allocation to more refined and optimization-based designs such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and convex-optimization-based methods [7-9]. While heuristic methods are low in design complexity, they cannot cope with the highly dynamic channel environment, thus leading to sub-optimal performance. On the upside, traditional optimization techniques can guarantee better results; however, the downside is that the complexity involved is quite high and that the centralized solution procedure suffers from scalability issues with an increase in the number of users, particularly in real-time problems.

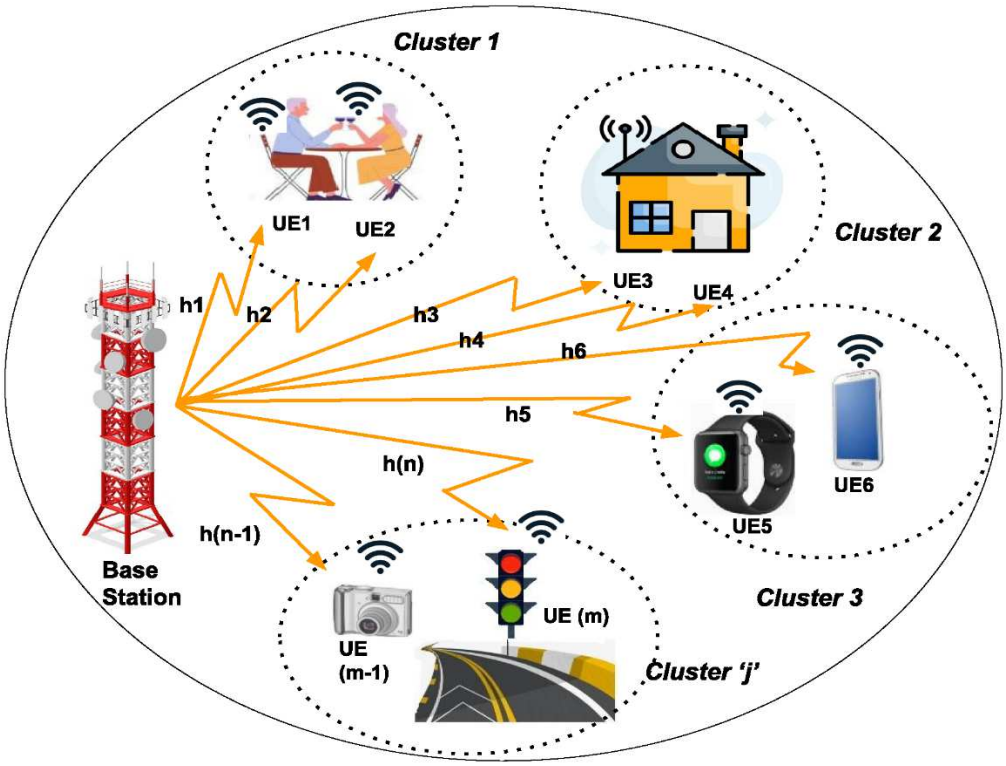


Fig. 1: Power domain allocation in downlink NOMA [4].

In recent times, Machine Learning (ML) and Reinforcement Learning (RL) techniques started being considered strong assets in solving complex resource allocation problems in wireless networks [10-13]. More specifically, DRL approaches such as Deep Q-Networks (DQN) have been promising for dynamic power allocation in NOMA systems. The advantage of these approaches is that they learn by interacting with the environment and by adapting their strategy toward a reward given as feedback. In a NOMA system, RL algorithms can dynamically learn the optimum power allocation policy that maximizes the system utility (e.g., sum rate, fairness, or energy efficiency) subject to the QoS constraint [14]. Despite these advances, most existing RL-based approaches rely on a centralized learning framework, where a single agent is responsible for making global decisions. While effective in certain scenarios, such centralized approaches may suffer from scalability issues, increased computational burden, and limited adaptability in highly dynamic or distributed network environments. To address these limitations, Multi-Agent Reinforcement Learning (MARL) has recently gained attention as a more scalable and flexible alternative. In a MARL framework, multiple agents independently learn and make decisions while interacting with each other and the environment. This decentralized learning paradigm enables better coordination among users, improved adaptability to local channel variations, and enhanced robustness in complex network conditions.

This research is motivated by the need to explore and compare various power allocation strategies, demonstrating especially how RL-based methods stand out against classical ones. Even though the literature points to several studies showing that RL can indeed improve NOMA performance, a comprehensive and fair comparison across multiple metrics like sum rate,

outage probability, bit error rate (BER), and fairness index seems to be missing. Furthermore, most RL- based approaches have been tested in simplified scenarios or without rigorous benchmarking against major classical methods such as GA, Fixed, or Equal power allocation. This makes the practical application in real-world scenario limited. As a result, a robust and decentralized approach such as the proposed MARL-based. The present article fills this void by analyzing the performance of five power allocations in detail within a 3-user downlink NOMA: Fixed, Equal, GA-based, RL-DQN-based, and MARL-based approach. The MARL algorithm serves to adjust power levels dynamically according to instant channel conditions, and the authors address this using adaptive and dynamic assignment through Monte Carlo simulations and assessing the performance of the system using sum rate, outage probability, BER, and Jain's index of fairness thus giving an overall assessment of each method's strengths and weaknesses.

Furthermore, a key challenge identified in some RL-based implementations is their occasional failure to outperform simpler methods under certain conditions. In order words, in some simulations, a scheme allocating power at a fixed or equal rate may perform better than RL-DQN because of sub-optimal training or an insufficient exploration of the action space [15]. To mitigate this, the RL-DQN framework is enhanced through modifications to the reward function and power allocation policy, encouraging the allocation of more power to users with weaker channel conditions. Additionally, a Signal-to-Interference-plus-Noise Ratio (SINR) enhancement factor is introduced during simulation to reflect improved decision-making capability [16]. Despite this, the needs to promotes fairness and reliability through cooperative learning among agents, leading to more balanced resource distribution and improved robustness, particularly under practical impairments such as imperfect SIC and channel estimation errors makes the performance of the RL-DQN worse. Based on this, the MARL is proposed to address the aforementioned challenges while maintain the advantage offers by the RL-DQN. From a practical standpoint, MARL-based power allocation deployment in real-world 5G systems will have to deal with training complexity, convergence speed, scalability, and generalization to unseen channel conditions. Hence, we contribute to this field by not only demonstrating the theoretical increase of MARL but also by proposing a lightweight implementation suitable for embedded systems and real-time applications.

The remaining parts of this paper are organized as follows: Section 2 reviews related works on NOMA power allocation strategies. Section 3 presents the system model and methodology for power allocation and metric computation. Section 4 provides the simulation setup, performance evaluation, and analysis of results. Finally, Section 5 concludes the paper and outlines future research directions.

2 Review of Related Works

Modern NOMA power allocation research has evolved rapidly due to the pressing need for high spectral efficiency, massive connectivity, and fairness in 5G and beyond. In this section, we look at the different techniques used in NOMA system by reviewing the power allocation problem, optimization techniques used, machine learning approach, fairness consideration, and imperfect SIC and practical constraints. Through this review, the gap in the current research was identified. Starting with power allocation in NOMA system, early researchers usually opted for fixed or equal power allocation schemes because they were easy to deal with and provided deterministic performance as shown in Fig. 2 [4-5]. With fixed allocation, users that experience poor channel conditions are allotted more power so as to carry out SIC, while power for users with better channel conditions is kept less [17]. It is simple to implement in practice but does not accommodate channel variations and thus can be at times very sub-optimal from a spectral efficiency perspective. Equal power allocation simply allocates power equally among users but fails to satisfy QoS requirements for the weaker users, thereby raising the probability of outages and unfairness [18-19]. To overcome the limitations of heuristic schemes, researchers turned to metaheuristic algorithms, such as PSO and GA, which iteratively search for near-optimal allocations [7-9]. For instance, the PSO-based approach for user pairing and power assignment of Masaracchia et al [20] gave considerable power savings while still maintaining QoS constraints in terrestrial and UAV-assisted NOMA systems [20]. Other works have used GA to achieve outage balancing and energy efficiency in heterogeneous NOMA networks, again highlighting the suitability of GA to non-convex, multi-objective problems [21]. Other approaches include classical optimization methods. Jain's index-based fairness optimization and max-min rate balancing usually involve convex or mixed-integer programming. But even if the solutions are theoretically sound, these methods have very heavy computational loads and become unsuitable in fast-varying environments.

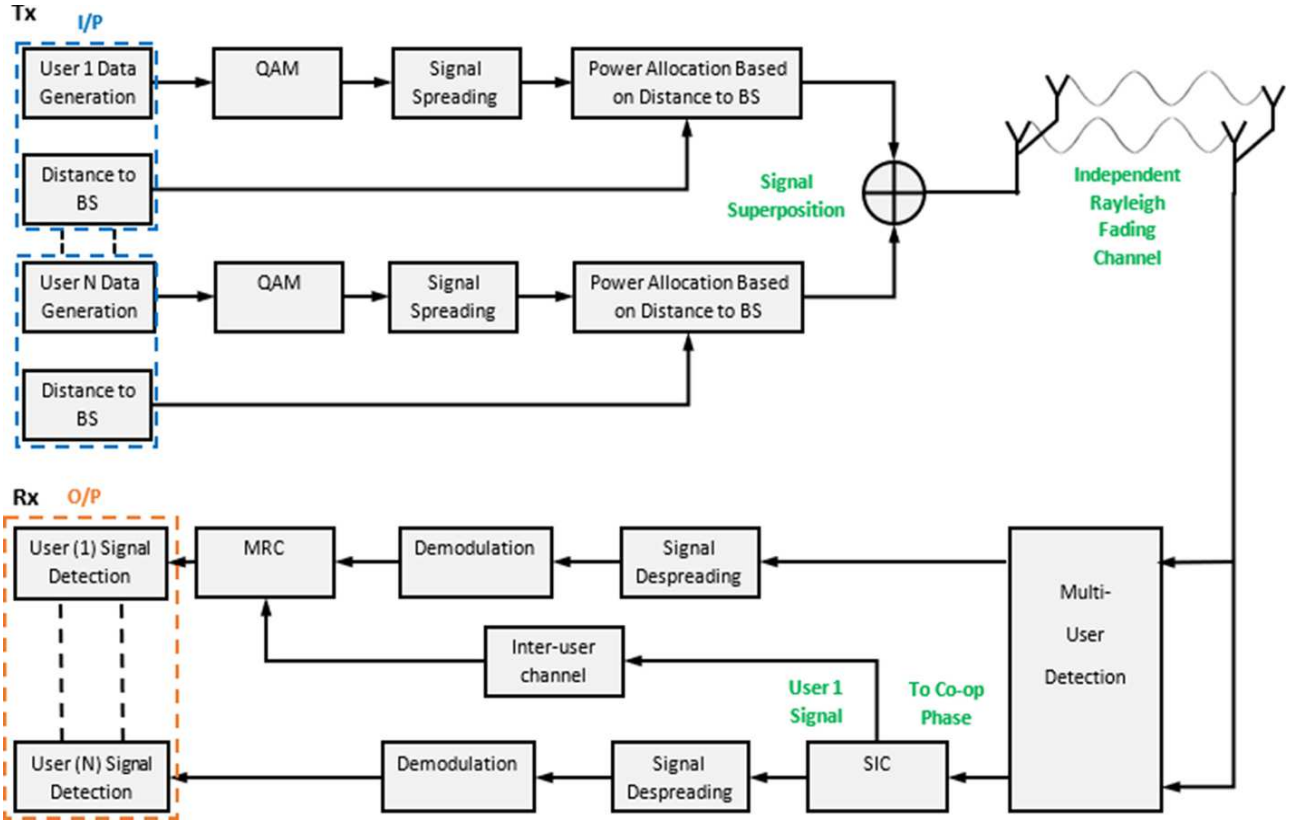


Fig. 2: NOMA system block diagram [5].

With the advancement in computational power and accessibility in wireless resource allocation, techniques such as the ML tools, RL, particularly Q-learning and DQN, have become attractive approaches for time-varying power allocation. In the area of Tabular Q-Learning, [9] applied Q-learning to joint channel prediction and power allocation in MISO-NOMA and found improvement over traditional methods, such as round-robin and equal power allocation. Scaling, however, is hindered by the dependency on discrete state-action representations. For Deep Learning–aided SIC and Outage Compensation, [22] proposed a Convolutional Neural Network (CNN) based SIC to improve decoding performance under imperfect cancellation, hence leading to more reliable allocation decisions. Likewise, a Deep Neural Network (DNN) based compensation was introduced in another piece to guarantee QoS during cell outages [11]. In the area of DQN and Actor-Critic Methods, [23] proposed a DRL, specifically DDPG and solved decentralized MIMO-NOMA power control in vehicular edge computing, reducing latency and maximizing long-term utility versus heuristic or centralized schemes. In addition, [13] put forth an SAC-based method for joint beamforming and power allocation in millimeter wave (mmWave) NOMA, having shown to perform well in realistic channel settings. Despite these advances, RL methods often struggled with unstable or inconsistent results particularly in fairness and BER metrics unless the reward functions were carefully designed. Simple DRL without fairness-conscious objectives may yield high throughput but poor equity across users.

In the area of Fairness, a greater focus has been exerted on ‘fairness' in NOMA with indices like Jain’s fairness index or min-max rate balancing being quoted in solutions, but very few have intermixed this with power control. An algorithm that was implemented in [24] provided an equitable method but, does not involve learning methodologies. In designing such models in which some so-called fairness is possible, even in the above, the impact is more significant, since some devices may have low power supply and may also relate to other investigations such as handling long delay duration devices in internet of things (IoT) network [11-15]. In addition, most existing literature assumes perfect SIC, which does not reflect real-world deployments. In practice, residual interference due to imperfect SIC leads to performance degradation, especially for users with lower power levels [25]. Works such as Islam et al [26] modeled SIC errors but did not propose adaptive solutions. Few studies have combined RL with robust power allocation under SIC uncertainty, making it a promising direction.

Having conducted a comprehensive review of the literature, it became apparent that some potential areas of interested were not addressed. Firstly, a majority of reinforcement learning studies pertain to minimizing sum rate. However, due to this wrong paradigm of trying to increase system capacity for a wireless network, most of them just concentrate on methods to minimize interference with no due regard for a reasonable access policy that can accommodate restrictive quality of service constraints.

In addition, the performance of adequate reward functions or appropriate approaches to initialization using heuristics may be a critical impediment to the performance of RL methods, with them being outperformed by more traditional methods, such as fixed and GA-based allocations [26-29]. Moreover, an obvious drawback refers to the limited number of those studies in which heuristic, optimization, and learning algorithms are presented and compared with various performance metrics such as outage probability and BER. Lastly, the closeness to reality in many case studies remains unacceptably low, as many scholars present overly simplified channel states instead of involving elaborate Monte Carlo methods or considering cases of non-ideal SIC.

Thus, the literature clearly shows the evolution of power allocation methods for NOMA from simple heuristics to advanced ML-driven schemes. While RL, particularly DRL, presents a promising framework, its practical deployment depends heavily on robust reward design, stability, and thorough benchmarking. Our work aims to contribute in this direction by proposing an MARL to address the challenges tin the current RL-DQN approach. A comprehensive comparison with the existing approaches will be carried out using performance evaluation metrics of sum rate, outage, BER, and fairness metrics. By uniting methodological rigor, comparative breadth, and practical realism, our study addresses a critical gap in the literature.

3. Research Methodology for The MARL NOMA System

In this section, the detailed system model and mathematical formulation for the proposed power allocation scheme in downlink NOMA systems, with an emphasis on the MARL are presented. The section is divided into the system model, channel model, optimization problem formulation, and our proposed hybrid algorithmic solution. A summary of the proposed methodology is as shown in Fig. 3.

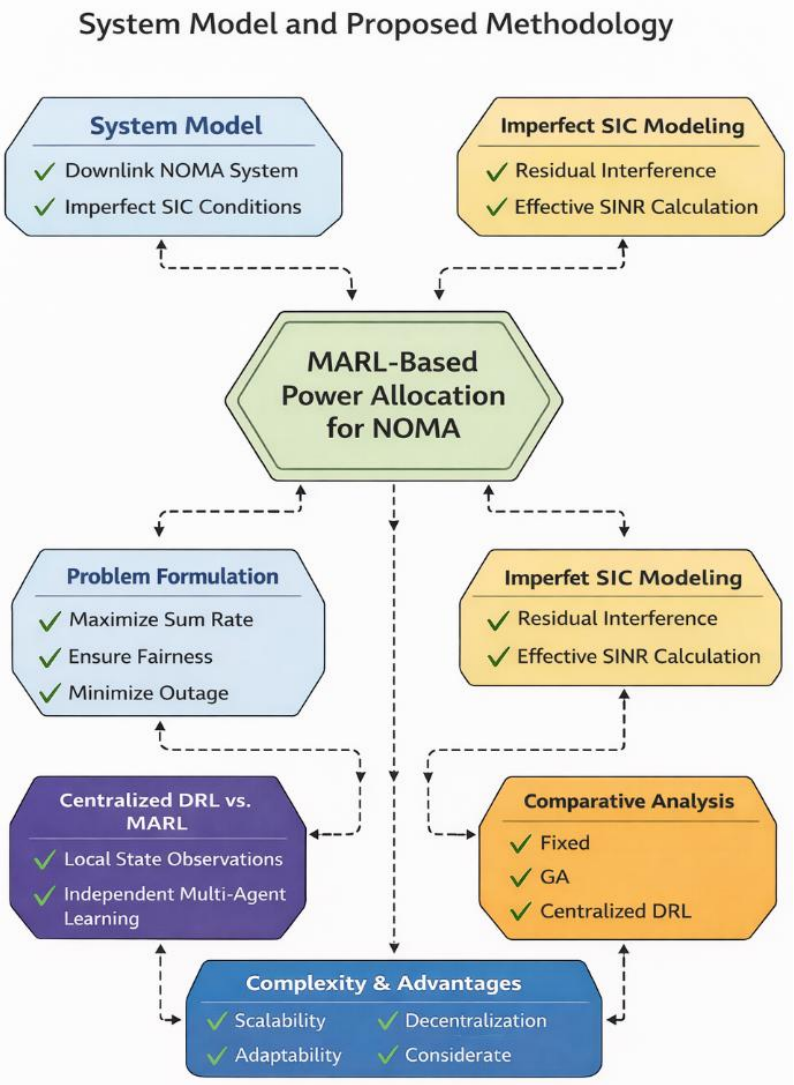


Fig 3. Proposed methodology for the MARL-based Power domain allocation in downlink NOMA.

3.1 NOMA System Model

Consider a downlink NOMA system comprising one base station (BS) and K users indexed by $k = 1,2, \dots, K$, each equipped with a single antenna. The base station is also assumed to be equipped with a single transmit antenna. The users experience different channel conditions and are sorted in descending order of their channel gains such that [30]

$$|h_1|^2 \geq |h_2|^2 \geq \cdots \geq |h_K|^2 \quad (1)$$

where h_K denotes the channel coefficient between the BS and the kth user.

Let the total transmission power at the BS be P_T, and the power allocated to the kth user be denoted by P_k, such that:

$$\sum_{k=1}^{K} P_k \leq P_T \quad (2)$$

Based on the above formulation, the transmitted signal becomes:

$$x = \sum_{k=1}^{K} \sqrt{P_k s_k} \quad (3)$$

where s_k defines the unit-power information signal.

In this NOMA system, the BS superposes the information signals intended for all users into a single composite signal, which is transmitted over the wireless channel. Each user applies SIC to decode its intended message.

3.2 Imperfect SIC Channel Modelling

The channel model considered is a Rayleigh fading channel with additive white Gaussian noise (AWGN). The received signal at user k is given by

$$y_k = h_k x + n_k = h_k x = \sum_{k=1}^{K} \sqrt{P_i s_i} + n_k \quad (4)$$

where x_i is the transmitted signal for the ith user, P_i is the power allocated to the user i, $n_k \sim \mathcal{CN}(0, \sigma^2)$ is the AWGN at user k, $h_k \sim \mathcal{CN}(0, \sigma^2)$ is the complex channel gain for user k.

Assuming perfect SIC, user k decodes and subtracts signals of users with worse channel gains (i.e.,$j > k$), treating signals of users with stronger channel gains (i.e.,$j < k$)as interference.
In practical systems, SIC is imperfect due to hardware impairments and channel estimation errors. To model this, a residual interference factor $\epsilon \in [0,1]$ is introduced. Based on that, the effective SINR at the user k is given by

$$\gamma_k = \frac{|h_k|^2 P_k}{\sum_{i=1}^{k-1} |h_k|^2 P_i + \epsilon \sum_{j=k+1}^{K} |h_k|^2 P_j + \sigma^2} \quad (5)$$

where P_i the allocated power, $\sigma^2$is the noise power, and $|h_k|^2$ is the channel gain for user k.

The above relation captures both the interference from stronger users and residual interference due to imperfect SIC.

3.3 Generalized Power Allocation Method

To successfully implement the power allocation method, we need to determine the SINR and the optimization formulation. For the SINR, using the signal model and the channel model, the SINR and achievable rate can then be determined based on equation (5). By performing the SIC on this equation, we get

$$R_k = \log_2 \left(1 + \frac{|h_k|^2 P_i}{\sum_{j-1}^{i-1} |h_k|^2 + P_j + \sigma^2}\right) \quad (6)$$

where R_k is the achievable rate of user k.

Based on this, the fairness between the users can then be measured using the Jain's fairness index given by

$$J = \frac{\sum_{k=1}^{K} R_k}{K \sum_{k=1}^{K} R_k^2} \quad (7)$$

The next stage is to formulate the power allocation as an optimization problem. To do this, the objective will be maximized the sum rate while ensuring fairness and minimizing outage probability and BER. Starting with the optimization problem, the first stage is to define the objective function by maximizing the weighted sum-rate using the relation:

$$\max_{P_k} \sum_{k=1}^{K} \omega_k R_k \quad subject\ to: \sum_{k=1}^{K} P_k \leq P_T, P_k \geq 0 \tag{8}$$

where ω_k is the weight reflecting user priorities or fairness constraints.

The objective is to determine the optimal power allocation that maximizes system performance while ensuring fairness and reliability. In addition, the updated R_k, fairness, and the outage probability are as follows:

$$R_k = \log_2(1 + \gamma_k) \tag{9}$$

$$J = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} R_k^2} \tag{10}$$

$$P_{out} = \mathbb{P}(R_k < R_k^{th}) \tag{11}$$

For the BER, a binary modulation scheme is used to determine the average BER for user k over Rayleigh fading channel given by

$$BER_k = \frac{1}{2}\left(1 - \sqrt{\frac{\gamma_k}{1+\gamma_k}}\right), with\ \gamma_k = SINR_k \tag{12}$$

Based on the above formulations, this is how the power allocation can be modeled as an optimization problem. Here, the main goal is to maximize the sum rate while ensuring fairness and minimizing outage probability and BER. And these formulations ensure that the set goal are achieved. Due to the coupled interference structure and QoS constraints, the problem is non-convex, motivating the use of learning-based approaches. In the next section, the proposed MARL and the RL-DQN based power allocation are presented in comparison with other techniques.

3.4 Conventional and RL-DQN Based Power Allocation

To solve the power allocation problem in comparison to conventional approaches, a RL-DQN that combines the global search capability DQN with the adaptive learning and decision-making strength of RL is used and are presented as follows:

a) Fixed and Equal Power Allocation: In fixed allocation, the power vector position $P = [P_1, P_2, \dots, P_K]$ is pre-defined such that lower-index users get higher power.
For equal allocation, the vector for the power allocation is given by

$$P_i = \frac{P_{total}}{K} \tag{13}$$

These formulations are then used to compute the fixed and the power allocation.

b) Genetic Algorithm: The concept GA is a concept that has gained wide popularity in optimization problems as it evokes the principle of natural evolution. It has been observed that GA are one of the most effective strategies in solving complex systems especially high-speed carrier communication technology (HPLC) which is not linear and concave, making it a critical tool also in energy allocation for NOMA systems which usually have very large complex non-convex problem associated [8]. In the perspective of our NOMA system, GA is introduced to optimize a pool of energy vectors iteratively to find the best or suboptimal solution by means of satisfying with precision the tradeoff between higher throughput to the system and providing fairness to the users. Every sequence (or chromosome) included in the latter embodies a probable dispersal of power signals in favor of the set of three consumers without impinging on the total power being utilized by the base station beyond its physical limit [31].

The algorithmic process summarized in Table 1 starts by establishing a population in which the power vectors are allocated randomly within the boundaries where each of the Power antennas P_i , that is within the boundaries given by

$$\sum_{i=1}^{K} P_i \leq P_{total} \quad (14)$$

Each individual is then evaluated using a fitness function that combines multiple performance metrics, such as sum rate and Jain's fairness index. A selection process follows, often using techniques like roulette wheel or tournament selection to choose the best-performing individuals to pass their "genes" to the next generation [32].

Crossover and mutation operators are applied to create offspring that inherit traits from the selected individuals while introducing variability. Crossover ensures exploration of new regions of the solution space by combining two parent solutions, whereas mutation helps avoid local minima by slightly altering a gene (i.e., power coefficient) with a small probability. The process repeats over several generations until a stopping criterion is met such as reaching a maximum number of generations or achieving convergence in fitness scores.

After applying the GA, the near-optimal power allocation vector can be determined as

$$\sum_{i=1}^{K} P_i \leq P_{total} \quad (15)$$

where P_i^* is the optimized power allocated to user i.

Equation (14) can then be updated to become

$$\sum_{i=1}^{K} P_i^* \leq P_{total} \quad (16)$$

Table 1: GA Algorithm development.

Algorithm GA
Input: Raw Data of the P
1. Initialize a population of candidate power allocation vectors.
2. Evaluate each individual using the defined fitness function.
3. Select parents based on fitness scores.
4. Apply crossover to generate offspring.
5. Apply mutation to introduce diversity.
6 Update the population with new individuals.
7. Repeat steps 2–6 until convergence or maximum generations are reached.
Output: P_i^*, J

c) RL-DQN Approach: The RL-DQN uses a neural network to approximate the Q-value function and select optimal power allocation actions. The main key components are the state, action and the reward. In the model, the technique used for the power allocation utilizes the MDP. The technique is as shown in Fig. 4. From the figure, the flowchart outlines the main steps involved in applying RL-DQN for power allocation in NOMA systems. The process begins by setting up a neural network designed to estimate the expected rewards for different actions in a given system state. In this context, the "state" might include information such as channel conditions, current power levels, and previous allocation decisions. Based on the observed state, the agent typically a base station chooses an action using an ε-greedy strategy, which means it either selects the best-known option or occasionally explores other possibilities to gather more experience [33]. Once an action is taken, such as allocating a specific amount of power to a user, the system responds with a new state and a reward that reflects how good that decision was perhaps in terms of improved data rates, reduced outages, or better user fairness. These interactions are stored in memory and used to

gradually improve the neural network's predictions by learning from past outcomes. Over time, this learning loop helps the system make more effective power allocation decisions, especially in dynamic and unpredictable wireless environments.

Based on this, the MDP uses the state s including the user channel gains and also, the previous power allocations. The action is a set of power values given by $\{P_k\}$. The reward r is computed using the utility function given by

$$r = \alpha \sum R_k - \beta \sum P_k - \gamma \sum P_{out,k} \tag{17}$$

where $\alpha, \beta, and\ \gamma$ are tuning parameters.

Based on the above formulation, the Q-network is then trained by utilizing the Bellman equation given by

$$Q(s,a) = r + \delta \max_{a'} Q(s',a') \tag{18}$$

where δ is the discount factor.

In addition, the DQN is trained to minimize the mean-squared loss using the relation:

$$Q(s,a) = r + \delta \max_{a'} Q(s',a') \tag{19}$$

By integrating RL-DQN, this enables adaptive and near-optimal power allocation that dynamically responds to varying channel conditions. In the end, to test the performance of this technique, the metric measurements of sum rate, outage probability, BER, and fairness index will be used. These metrics provide comprehensive insight into the system's efficiency, reliability, and user equity. Despite this advancement, the RL- DQN approximates the action-value function and it suffers from high dimensional action space, scalability limitations, and centralized computational burden.

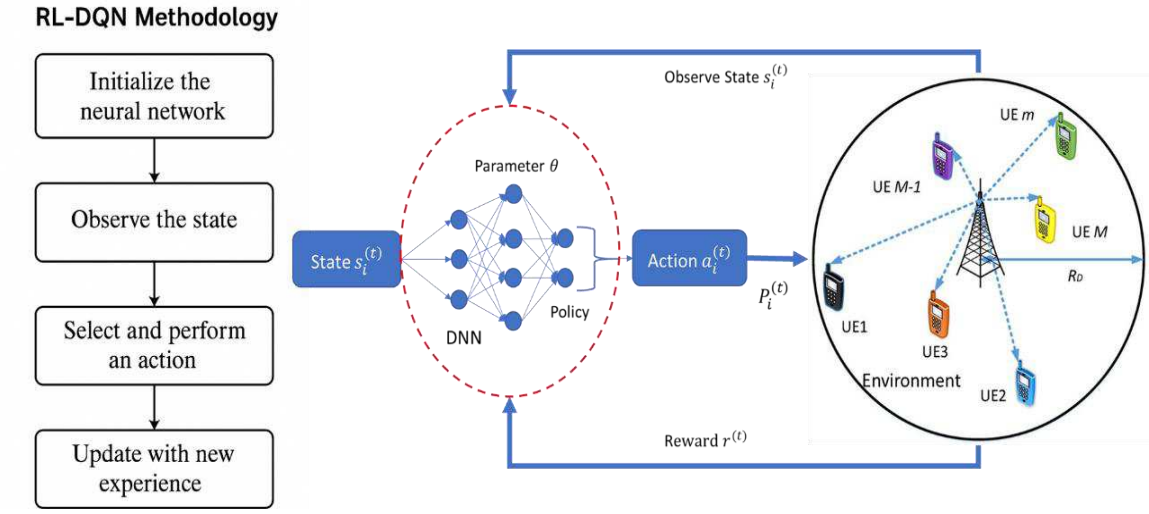


Fig. 4 The RL-DQN approach.

3.5 Proposed MARL Based Power Allocation Approach

With the limitations of the RL-DQN as presented, we propose a decentralized MARL framework in which each user acts as an independent learning agent. This is done to enhance scalability and adaptability by extending it to a multi-agent setting, where each user is modeled as an independent learning agent. Unlike the RL-DQN approach, where a single agent determines the power allocation vector, the MARL formulation enables distributed decision-making. To formulate this, each agent k observes a local state given by

$$s_k = \{|h_k|^2, I_k, P_k^{t-1}\} \tag{20}$$

where I_k defines the interference which was experienced by user k. This reduces dimensionality compared to centralized state representation.

Based on the above formulation, the action of each agent now corresponds to choosing its own power coefficient which can be found using

$$a_k \in \mathcal{A}_k \tag{21}$$

Subject to a global power constraint given by

$$P_k = \frac{a_k}{\sum_{i=1}^{K} a_i} P_T \tag{22}$$

Equation (22) ensures feasibility and coordination among agents. To ensure coordination among agents, a cooperative reward design based on a shared global reward that can be computed using equation (17) as well. This encourages the agents to cooperate rather than compete, maintain fairness, and avoid bias power allocation.

The agents now employ an independent DQN (or parameter sharing scheme) to learn its policy. This now leads to updated Bellman equation in (18) to becomes

$$Q_k(s_k, a_k) \leftarrow r + \delta \max_{a_k'} Q_k(s_k', a_k') \tag{23}$$

This decentralized learning approach reduces computational burden at the base station and improves scalability in dense networks. The summary of the MARL is as shown in Fig. 5 and the algorithm process is as shown in Table 2.

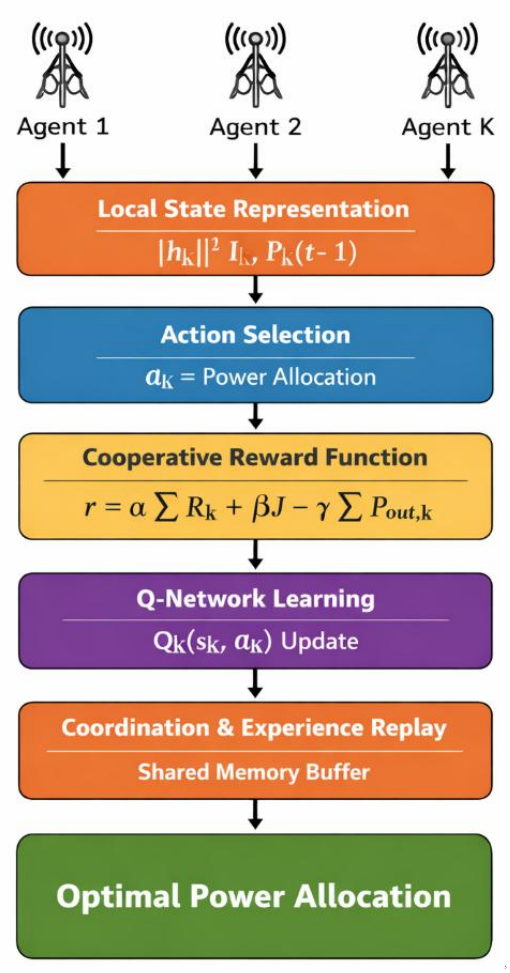


Fig. 5 The proposed MARL approach.

Table 2: MARL Algorithm development.

Algorithm: MARL-Based Power Allocation for Downlink NOMA
Input: Number of users K, total transmit power P_T, learning rate α, discount factor δ, exploration rate ϵ, number of episodes N_{ep}
Step 1: Initialization
1. Initialize DQN parameters θ_k and target networks θ_k^* for each agent $k = 1,2,\dots,K$ 2. Initialize replay memory $\mathcal{D}_k$ for each agent 3. Define discrete action space $\mathcal{A}_k$
Step 2: Training Phase
4. For episode = 1 to N_{ep} do 5. Generate channel gains $\{h_k\}_{k=1}^K$ 6. Observe initial states s_k
Step 3: Interaction Loop
7. For each time step t do 8. For each agent k: 9. Select action a_k using ε-greedy policy 10. End for
Step 4: Power Normalization
11. Compute: $P_k = \frac{a_k}{\sum_{i=1}^K a_i} P_T$
Step 5: Environment Interaction
12. Compute SINR γ_k under imperfect SIC 13. Compute metrics: sum rate, fairness, outage 14. Compute global reward: $r = \alpha \sum R_k - \beta \sum P_k - \gamma \sum P_{out,k}$
Step 6: Learning Update
15. For each agent k: 16. Observe next state s_k' 17. Store transition (s_k, a_k, r, s_k') 18. Sample mini-batch 19. Update Q-network: $Q_k(s_k, a_k) \leftarrow r + \delta \max_{a_k'} Q_k(s_k', a_k')$ 20. End for
Step 7: Target Network Update
21. Update: $\theta_k^* \leftarrow \theta_k$
Step 8: Inference Phase
22. Select optimal action: $a_k = argmax Q_k(s_k, a_k)$ 23. Compute final power allocation P_k
Output: Optimal power allocation policy π_k^* for each agent k

4 Results and Discussion

In this section, the results and discussion of the proposed MARL-based power allocation framework under various channel conditions are presented. Starting from the simulation parameters, the setup for the simulation is outlined. We then look at the simulation results using the performance metrics, including sum rate, outage probability, fairness index, and BER evaluated using Monte Carlo simulations averaging over 10,000 iterations. The performance is benchmarked against conventional schemes, including fixed, equal, and GA-based power allocation methods, as well as the centralized RL-DQN approach.

4.1 Simulation Parameters

The simulation considers a downlink NOMA system with three users experiencing independent Rayleigh fading channels. The total transmit power is normalized to 1 W, and the noise power is set to 10^{-9} Watts. The remaining simulation

parameters are shown in Table 3 and were chosen based on a typical NOMA system for 5G network. Moreover, the users are sorted based on their channel gains, and the NOMA principle of SIC was applied. As can be seen from the table, the channel is modeled to be quasi-static, such that it remains constant for the given period of each transmission block. In terms of the learning-based approaches, both centralized RL-DQN and MARL agents are trained over 500 episodes using a learning rate of 0.001 and a discount factor of 0.95 for fair comparison. In the MARL framework, each user becomes independent agent and operates separately with local observations, while a shared global reward ensures cooperative behavior.

Table 3: Simulation parameters of the NOMA system.

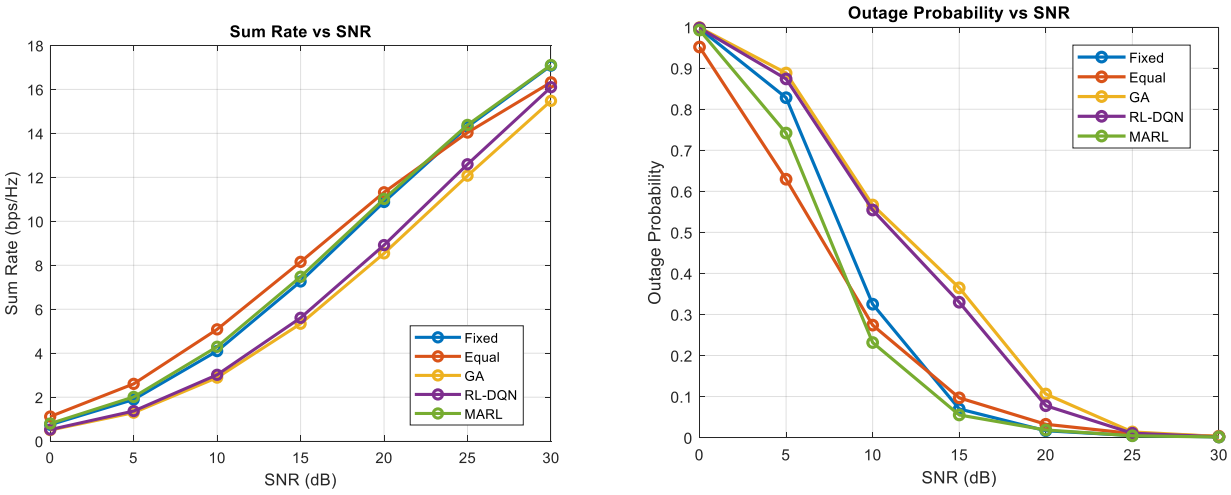
Parameter	**Value**
Number of users	3
Total transmit power (P)	1 W
Noise power (σ^2)	10^{-9} W
Modulation	QPSK
Channel model	Rayleigh fading
Iterations (Monte Carlo)	10,000
MARL/RL-DQN episodes	500
MARL/RL learning rate	0.001
Discount factor (γ)	0.95

4.2 Results for The Metric Performance of the NOMA System

In the coming subsections, the results for the core performance evaluation of the proposed power allocation strategies for a NOMA-enabled 5G system are presented using the four key performance metrics of sum rate, outage probability, fairness index, and BER under different SNR conditions. For the comparison of the methods, we used the Fixed, Equal, GA, RL-DQN, and the proposed MARL approach. Each metric provides insight into different aspects of the system's reliability, efficiency, and user experience.

4.3 Results for Sum Rate Performance

The sum rate is a primary indicator of spectral efficiency in NOMA systems. Using the model in section III, the simulation was run and the results for the different approaches are shown in Fig. 6 (a). From the figure, it can be seen that the sum rate performance of the different power allocation strategies is illustrated as a function of the SNR. Looking at the results, the centralized RL-DQN approach offers a good spectral efficiency, especially in the moderate-to-high SNR region, with regards to its global view of the system state. However, the proposed MARL framework demonstrates competitive performance, closely approaching that of the fixed and equal cases that are usually not fair in channel assignment. This performance is attributed to the ability of MARL agents to learn cooperative strategies through a shared reward structure, despite relying on local observations. In addition, the conventional schemes of Fixed and Equal allocation schemes lack adaptability which leads to suboptimal resource utilization, especially when user channel conditions are highly asymmetric. For the GA-based approach, it shows an improved performance but still lower in terms of the total sum rate as the channel quality improves. This is because its performance remains limited by its iterative nature and lack of real-time adaptability. Compared to GA, both RL-DQN and MARL benefit from learning-based generalization, enabling faster and more efficient decision-making after training. With regards to the performance gap between MARL and centralized RL-DQN, this can be attributed by the slight increase at very high SNR levels, where global optimization provides marginal advantages. This finding shows that while centralized learning may offer peak performance, MARL provides a scalable alternative with competitive efficiency.



(a) (b)

Fig. 6 Simulation results for a) Sum rate against SNR. b) The outage probability against SNR.

4.4 Results for Outage Probability

The outage probability is a key reliability metric that reflects the proportion of users unable to achieve the minimum required data rate. The results for the outage probability are as shown in Fig. 6 (b). The results provide a clear insight into how reliably each power allocation method maintains a minimum required data rate under varying channel conditions. A lower outage probability implies higher reliability, a critical requirement for modern wireless systems, particularly in mission-critical applications. Also, it is evident that all strategies benefit from increased SNR, with outage probability decreasing significantly as signal strength improves. From the results, at the low-to-moderate SNR regions, where efficient power allocation is critical for maintaining user connectivity, it was observed that the MARL offers a lower outage compared to the GA and RL-DQN and also achieved comparable results with the conventional techniques. This shows the MARL adaptability even at low SNR condition (below 5 dB). Thus, it can be said that the MARL framework effectively reduces outage probability by enabling agents to cooperatively prioritize weaker users through the shared reward mechanism. At 10 dB, the MARL reduces outage probability to below 10%, and this shows its adaptive capability to allocate resources effectively, even under poor channel conditions. The GA-based method also achieves relatively low outage probability but still worse than the Fixed and Equal schemes. However, its performance is less consistent across different SNR levels due to its stochastic optimization process. At higher SNR levels, the performance differences among the schemes become less pronounced, as most users are able to meet their rate requirements regardless of the allocation strategy. These results indicate that the primary benefit of learning-based approaches, including the MARL, lies in enhancing robustness under constrained channel conditions.

4.5 Results for Fairness Index

While high throughput is essential, fairness among users remains equally critical in multi-user scenarios of NOMA system. Using the fairness index formulation in section III, the simulation was carried out and the results are shown in Fig. 7 (a). From the figure, it was observed that the MARL approach maintains high fairness levels across all SNR values compared to the RL-DQN approaches while having a comparable result with that of fixed allocation. This strong fairness performance of the MARL can be attributed mainly to its cooperative reward design, and this encourages agents to consider global system objectives to improve the system. In addition, it enables a balanced resources allocation even in decentralized settings. Unlike Equal allocation, that enforces uniform power distribution without considering channel disparities, the MARL approach shows that by dynamically adapting power levels based on local channel conditions, the fairness of each user can be preserved. The GA-based approach also provides reasonable fairness performance; however, it exhibits fluctuations due to its stochastic nature. The findings show the effectiveness and the need to incorporate fairness-aware objectives into decentralized learning frameworks for multi-user systems.

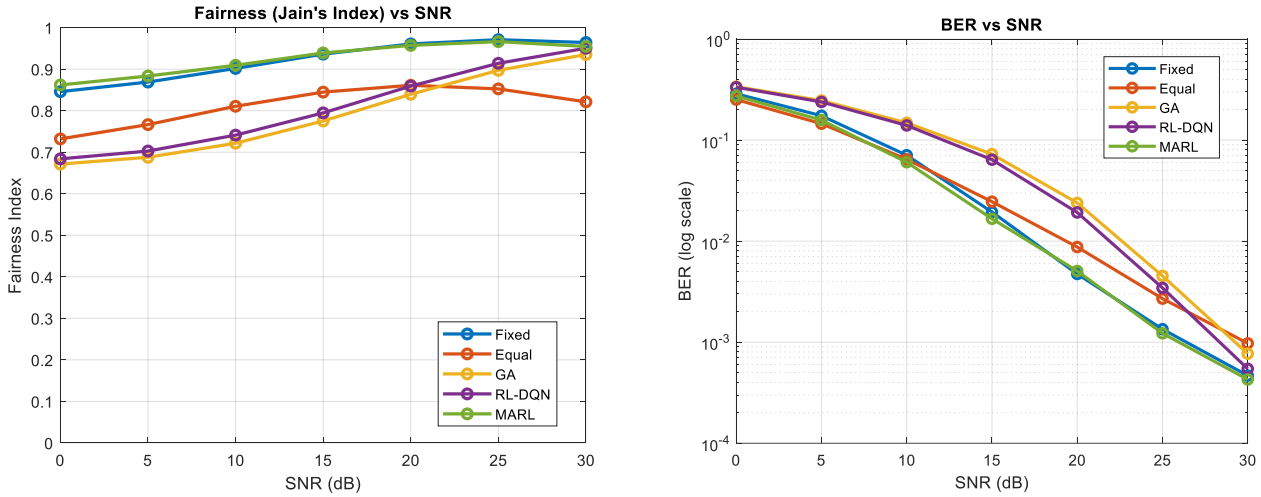


Fig. 7 Simulation results for a) The fairness index against SNR. b) The BER against SNR.

4.6 Results for Bit Error Rate (BER)

While The BER measures how accurately each method transmits data without errors and this is vital for high-quality communication since it provides an insight into the reliability and accuracy of data transmission. Based on the simulation, the BER results are shown in Fig. 7 (b). From the results, at low SNR region (0 to 10 dB), it can be seen that the conventional approaches and the MARL have comparable results while RL-DQN and GA have worse BER. The improved BER performance of MARL can be attributed to its ability to manage interference effectively through adaptive and cooperative power allocation strategies. For the GA approach, The GA method also demonstrates improved BER performance; however, its gains are less consistent due to its iterative optimization nature. As the SNR increases, the BER of both approaches decreases and this is expected since they are both operating at a better condition. Even with that, the MARL significantly lowers BER below $\mathbf{10^{-3}}$ has a better performance compared to other techniques showing it adaptability and advantage of decentralization affirming its uniqueness in maintaining data integrity even at very high transmission rates. It is worth noting that BER performance is closely linked to the effectiveness of SIC. Therefore, the improved BER observed in MARL indicates its ability to account for residual interference effects introduced by imperfect SIC, even under decentralized decision-making. Thus, it can be concluded that the log-scale BER plot clearly underscores the inherent strength of the proposed MARL method. These outcomes indicate that MARL's policy optimization, driven by real-time feedback and continuous learning, effectively minimizes errors by dynamically distributing power where it's most beneficial demonstrating an ideal balance of efficiency, reliability, and resource optimization.

4.7 Impact of Imperfect Channel State Information

To further evaluate the robustness of the proposed framework under practical conditions, the performance of both approaches was analyzed under in the presence of imperfect channel state information (CSI) to account for inaccuracies caused by channel estimation, feedback delay, or quantization [34-36]. The CSI was modeled by taking into consideration the AWGN to add realistic inaccuracies in channel acquisition. Thereafter, the simulation was run using the same setup and the results are shown in Fig. 8. From the results, it was observed that both approaches experience a degradation in sum rate performance as compared to the perfect CSI scenario, which is expected due to the mismatch between estimated and actual channel conditions. However, the extent of this degradation differs between the different approaches. For the MARL, because of its adaptability, it can be seen it has a higher sum rate at low SNR condition (0 to 10 dB) compared to the other approaches. However, as the SNR increases, the conventional approaches achieved better sum rate which is expected as they don't have fair assignment to the users as some may appear stronger the others. This is clearly seen by looking at the fairness results as they index decreases while that of the MARL achieved an index as high as 0.9. in addition, looking at the outage probability results, the MARL still outperformed other approaches as the outage approaches 0 as the SNR increases. This was also supported by the BER performance results. Thus, it can be said in comparing the two best approaches of the RL-DQN and MARL, the centralized RL-DQN approach maintains slightly higher sum rate performance across all SNR values, owing to its more aggressive power allocation strategy. However, this benefit is reduced in the imperfect CSI settings, where the centralized policy is more affected by the errors of estimation because of the dependence on the global channel knowledge. Conversely, the performance trend of the proposed MARL framework is more stable. Its sum rate is slightly lower than that of the RL-DQN as the SNR increases, but the deterioration caused by CSI imperfections is reduced. This is due to the fact that the MARL is decentralized, meaning that

decisions are made based on local observations, which naturally causes the system to rely less on accurate information of global channels.

Moreover, the strength of the MARL in imperfect CSI reflects the applicability of this algorithm to realistic wireless networks, where channel estimation inaccuracies are inevitable. The findings suggest that the MARL provides an optimal trade-off between the performance and the robustness, especially in the case of unreliable or noisy channel information. Overall, these results support the practicality of the suggested MARL framework and indicate that it cannot only serve as a competitive framework in optimal conditions but also can ensure its stable functioning in more realistic and demanding communication conditions.

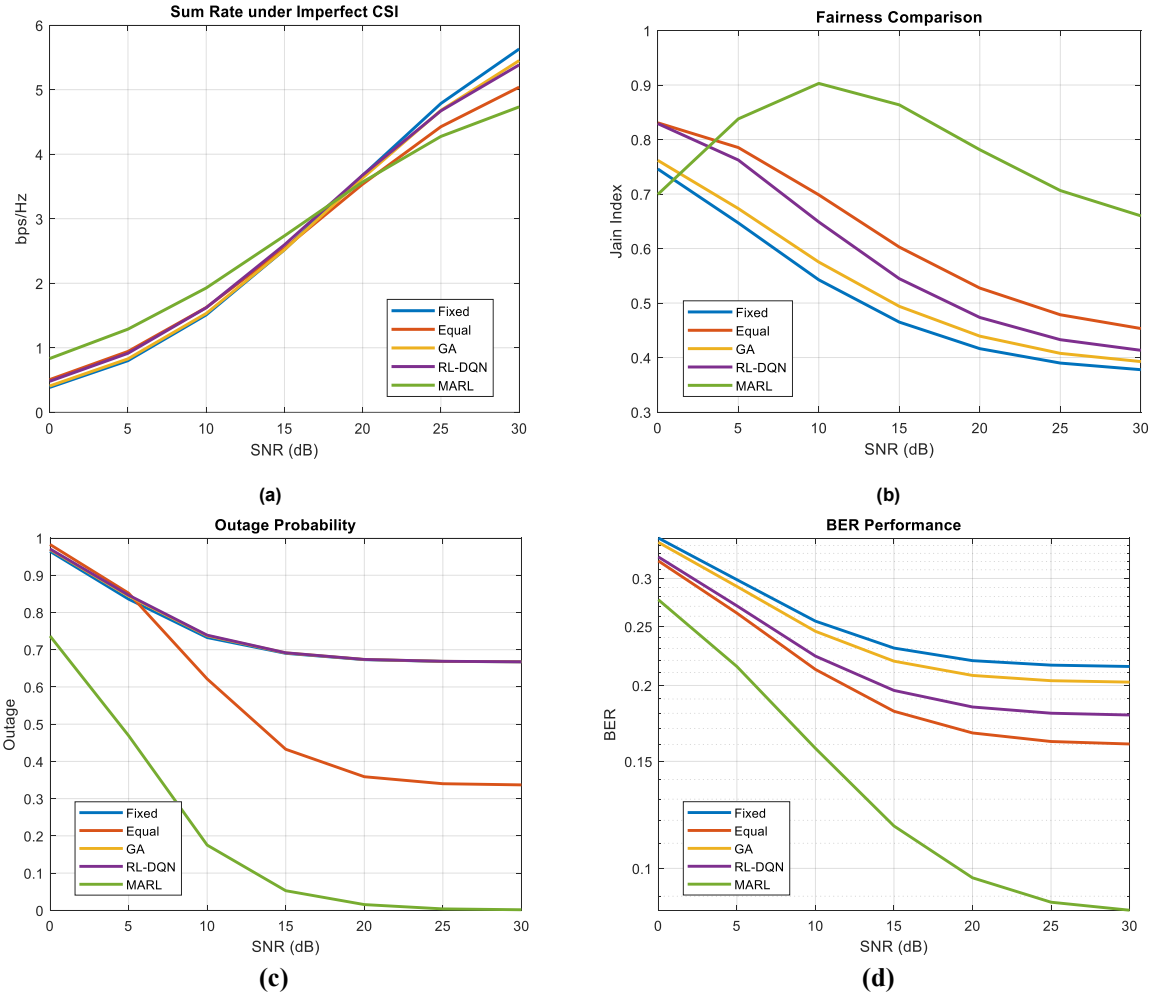


Fig. 8 Simulation results for the impact of Imperfect CSI a) Sum rate against SNR, b) Fairness, c) Outage probability, and d) BER.

5 Conclusion

In conclusion, this paper proposed a MARL-based power allocation framework for downlink NOMA systems in 5G communication networks. The proposed approach was evaluated against conventional and intelligent baseline schemes, including fixed, equal, GA, and RL-DQN-based power allocation methods. All schemes were assessed through extensive Monte Carlo simulations with 10,000 iterations, considering key performance metrics such as sum rate, outage probability, fairness index, and BER. The simulation results demonstrate that the proposed MARL framework outperforms conventional approaches, including fixed, equal, and GA-based schemes, across all considered performance metrics. More importantly, when compared with the RL-DQN approach, a clear performance trade-off is observed. While RL-DQN achieves the highest sum rate due to its throughput-oriented optimization as the SNR increases, the MARL framework provides superior performance in terms of fairness, outage probability, and BER. This improvement is primarily attributed to the decentralized and cooperative nature of MARL, which enables agents to balance resource allocation more effectively among users, particularly in heterogeneous channel conditions.

Furthermore, the incorporation of imperfect SIC and channel estimation errors (imperfect CSI) provides a more realistic evaluation environment. Under these practical impairments, the MARL approach demonstrates enhanced robustness,

maintaining stable performance with reduced sensitivity to channel uncertainties. This highlights its suitability for real-world deployments where perfect channel knowledge cannot be guaranteed. Overall, the findings indicate that while RL-DQN is effective in maximizing spectral efficiency, the proposed MARL framework offers a more balanced and reliable solution by jointly optimizing system fairness, reliability, and error performance. This makes the MARL a strong candidate for intelligent radio resource management in next-generation wireless networks.

As a future direction, it is still significant to apply the proposed framework to high-mobility cases and networks of large scale. Also, it would be possible to incorporate the latest learning paradigms, like actor-critic, federated learning, and meta-learning, to increase the adaptability and scalability of the system. Lastly, software-defined radio (SDR) platform implementation in real-time and experimental validation via field trials would be of considerable use in gaining insight into real-world deployment challenges and performance.

Acknowledgements

The authors extend their appreciation to the image processing group, Department of Electrical, Telecommunications and Computer Engineering, Kampala International University, Uganda

Authors contributions

The Conceptualization, framework, and methodology were done by MDB. MDB and BOS developed the mathematical and simulation modelling of the NOMA MARL, while OFC, ABK, and EE conducted the validation of the algorithms. Formal analysis was done by MDB, ABK, BOS, OFC, and EE. OFC and EE conducted techniques comparison and investigation, MDB gathered the required resources, and BOS, OFC, ABK, and EE participated in the original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding

No funding was received for this research

References

[1] Z. Ding, F. Adachi, and H. V. Poor, "The Application of MIMO to Non-Orthogonal Multiple Access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[2] S. P. Indraganti et al., "Survey of Cooperative NOMA for Beyond 5G: State-of-the-Art, Applications and Research Directions," in IEEE Open Journal of the Communications Society, vol. 7, pp. 40-94, 2026, doi: 10.1109/OJCOMS.2025.3648469.

[3] Z. Xiao, L. Zhu, J. Choi, P. Xia and X. -G. Xia, "Joint Power Allocation and Beamforming for Non-Orthogonal Multiple Access (NOMA) in 5G Millimeter Wave Communications," in IEEE Transactions on Wireless Communications, vol. 17, no. 5, pp. 2961-2974, May 2018, doi: 10.1109/TWC.2018.2804953.

[4] Hamedoon, S. M., Chattha, J. N., & Bilal, M. (2024). A novel user clustering and efficient resource allocation in non-orthogonal mutliple access for IoT networks. PLOS ONE, 19(9), e0309695. https://doi.org/10.1371/journal.pone.0309695.

[5] Elsaraf, Z., Ahmed, A., Khan, F.A. et al. Cooperative non-orthogonal multiple access for wireless communication networks by exploiting the EXIT chart analysis. J Wireless Com Network 2021, 79 (2021). https://doi.org/10.1186/s13638-021-01961-z

[6] M. Temiz et al., "Deep-Learning-Based Techniques for Integrated Sensing and Communication Systems: State-of-the-Art, Challenges, and Opportunities," in IEEE Open Journal of the Communications Society, vol. 6, pp. 5940-5968, 2025, doi: 10.1109/OJCOMS.2025.3586560.

[7] S. Lingyang, L. Yonghui, D. Zhiguo, and H. Vincent, “Resource Management in Non-orthogonal Multiple Access Systems: State of the Art and Research Challenges”, 2019. 10.48550/arXiv.1610.09465.

[8] M. B. Dadi and B. C. Rehaimi, "Performance of downlink MIMO-NOMA system in 5G Networks," *2023 IEEE International Workshop on Mechatronic Systems Supervision (IW_MSS)*, Hammamet, Tunisia, 2023, pp. 1-4, doi: 10.1109/IW_MSS59200.2023.10368947.

[9] P. Thakre, S. B. Pokle, R. Patel, N. Shrimankar and M. Dubey, "Performance of Deep Embedded Clustering and Low-Complexity Greedy Algorithm for Power Allocation in NOMA-Based 5G Communication," *2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP)*, New Delhi, India, 2024, pp. 1-6, doi: 10.1109/ICECSP61809.2024.10698406.

[10] G. M. Nguyen and J. -H. Yun, "Energy-Efficient NOMA Resource Allocation via Deep Transfer Reinforcement Learning," *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of, 2024, pp. 1875-1876, doi: 10.1109/ICTC62082.2024.10827109.

[11] W. Zhou, R. Zhang, G. Chen and W. Wu, "Integrated Sensing and Communication Waveform Design: A Survey," in IEEE Open Journal of the Communications Society, vol. 3, pp. 1930-1949, 2022, doi: 10.1109/OJCOMS.2022.3215683.

[12] T. P. Truong *et al*., "NOMA-Enhanced Intelligent Semantic Communication Networks using Deep Reinforcement Learning," *2024 International Conference on Information Networking (ICOIN)*, Ho Chi Minh City, Vietnam, 2024, pp. 426-428, doi: 10.1109/ICOIN59985.2024.10572204.

[13] W. Jiang, B. Han, M. A. Habibi and H. D. Schotten, "The Road Towards 6G: A Comprehensive Survey," in IEEE Open Journal of the Communications Society, vol. 2, pp. 334-366, 2021, doi: 10.1109/OJCOMS.2021.3057679.

[14] N. T. Yen Linh, P. N. Son and V. N. Quoc Bao, "Power Allocation for the STAR-RIS-NOMA Short Packet Communication System Using Deep Reinforcement Learning," *2024 International Conference on Advanced Technologies for Communications (ATC)*, Ho Chi Minh City, Vietnam, 2024, pp. 221-226, doi: 10.1109/ATC63255.2024.10908162.

[15] M. Liu, Y. Meng and Z. Zhang, "Deep Reinforcement Learning for Resource Allocation With Mixed Traffic in NOMA System," *2023 IEEE/CIC International Conference on Communications in China (ICCC)*, Dalian, China, 2023, pp. 1-2, doi: 10.1109/ICCC57788.2023.10233654.

[16] J. He, Z. Liu, Y. Zhang, Z. Jin and Q. Zhang, "Power Allocation Based on Federated Multiagent Deep Reinforcement Learning for NOMA Maritime Networks," in *IEEE Internet of Things Journal*, vol. 12, no. 9, pp. 12869-12884, 1 May1, 2025, doi: 10.1109/JIOT.2024.352290.

[17] G. M. Nguyen and J. -H. Yun, "Energy-Efficient NOMA Resource Allocation via Deep Transfer Reinforcement Learning," *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of, 2024, pp. 1875-1876, doi: 10.1109/ICTC62082.2024.10827109.

[18] M. K. Hasan, M. Shahjalal, M. M. Islam, M. M. Alam, M. F. Ahmed and Y. M. Jang, "The Role of Deep Learning in NOMA for 5G and Beyond Communications," *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Fukuoka, Japan, 2020, pp. 303-307, doi: 10.1109/ICAIIC48513.2020.9065219.

[19] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[20] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal Multiple Access for 5G and Beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[21] H. Zhang, Y. Xiao, S. Li, and L. Wang, "Optimization of Power Allocation in NOMA: State-of-the-Art and Future Directions," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 66–73, June 2019.

[22] J. Wang, H. Zhang, H. V. Poor, and L. Hanzo, "Machine learning aided sensing, edge computing and secure communications in intelligent 6G networks," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 18-25, Jun. 2021.

[23] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347-2381, Dec. 2017.

[24] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Wang, "Dynamic power allocation for NOMA systems: A reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1215-1227, Jun. 2019.

[25] M. M. Azari, F. Rosas, and S. Pollin, "Ultra Reliable UAV Communication Using Altitude and Cooperation Diversity," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 376–389, Jan. 2019.

[26] M. Chen, Y. Hao, L. Hu, M. S. Hossain, and A. Ghoneim, "Edge-Cooperative NOMA for Internet of Things: A Deep Reinforcement Learning Approach," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3561-3571, Aug. 2018.

[27] I. Ahmad, A. Ullah and W. Choi, "WiFi-Based Human Sensing With Deep Learning: Recent Advances, Challenges, and Opportunities," in IEEE Open Journal of the Communications Society, vol. 5, pp. 3595-3623, 2024, doi: 10.1109/OJCOMS.2024.3411529.

[28] J. Choi, "Power allocation for max-min fairness in NOMA systems," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055-2058, Oct. 2016.

[29] M. Alsenwi, N. H. Tran, T. M. Ho, and C. S. Hong, "eURLLC-enabled NOMA in 5G networks: Deep reinforcement learning for resource management," *IEEE Access*, vol. 7, pp. 181061-181073, 2019.

[30] H. Ye, G. Y. Li, and B. H. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163-3173, Apr. 2019.

[31] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077-1091, Mar. 2017.

[32] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721-742, 2nd Quart., 2017.

[33] W. Khalid, "DRL-Based Approach for RIS-Aided NOMA System in Short Packet Communications," 2025 8th International Conference on Circuits, Systems and Simulation (ICCSS), Ho Chi Minh City, Vietnam, 2025, pp. 90-93, doi: 10.1109/ICCSS65911.2025.11081794.

[34] S. Yan, K. -C. Chou, H. -H. Chen and Q. Guo, "Real-Time Downlink Resource Allocation in NOMA Systems: A Reinforcement Learning Approach," in IEEE Transactions on Vehicular Technology, vol. 74, no. 11, pp. 17779-17795, Nov. 2025, doi: 10.1109/TVT.2025.3575838.

[35] S. Ahmed et al., "Multiagent Reinforcement Learning for Joint Spectrum and Energy Optimization in CR-NOMA Enabled Internet of Unmanned Agents," in IEEE Internet of Things Journal, vol. 12, no. 24, pp. 55283-55296, 15 Dec.15, 2025, doi: 10.1109/JIOT.2025.3624882.

[36] D. Brodimas, A. Birbas, D. Kapolos and S. Denazis, "Intent-Based Infrastructure and Service Orchestration Using Agentic-AI," in IEEE Open Journal of the Communications Society, vol. 6, pp. 7150-7168, 2025, doi: 10.1109/OJCOMS.2025.3600706.